E-STATS STATISTICAL METHODOLOGY - SAS

INTRODUCTION

The U.S. Census Bureau produces the *E-Stats* report to provide national estimates of e-commerce activity by kind of business for establishments engaged in manufacturing, merchant wholesale trade, retail trade, and selected service industries. Estimates in this report for Accommodation and Food Services industries are based on data from the Annual Retail Trade Survey (ARTS) and administrative records. All other service estimates in this report are based on data from the Service Annual Survey (SAS) and administrative records. The following URL contains more information on ARTS:

http://www.census.gov/eos/www/sm.html

SAS questionnaires are mailed to a probability sample that is periodically reselected from a universe of firms operating in the United States and having paid employees. The sample includes firms of all sizes and covers both taxable firms and firms exempt from Federal income taxes. The following URL contains more information on SAS:

http://www.census.gov/econ/www/servmenu.html#services

COVERAGE

The service estimates of revenue and e-commerce revenue contained in this report are summarized by kind-of-business classification based on the *1997 North American Industry Classification System* (NAICS). NAICS is a classification system that groups establishments into industries based on the activities in which they are primarily engaged. This industry classification system was developed as a result of a joint effort by statistical agencies in Canada, Mexico, and the United States so that common industry definitions would allow for comparability in statistics on business activity across North America.

Estimates in this report for Accommodation and Food Services (NAICS 72) are based on results from the Annual Retail Trade Survey (ARTS). All other service estimates in this report are based on results from the Service Annual Survey (SAS) and are presented for selected industries in the following NAICS sectors:

NAICS Sector	Title
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
56	Administrative and Support and Waste Management and Remediation Services
62	Health Care and Social Assistance

71	Arts, Entertainment, and Recreation
81	Other Services (except Public Administration)

The following URL contains detailed information about NAICS and provides a comparison of the SIC and NAICS systems:

http://www.census.gov/epcd/www/naics.html

DOLLAR VALUES

All dollar values presented are expressed in current dollars; that is, the estimates are not adjusted to a constant dollar series. Consequently, when comparing estimates to prior years, users also should consider price level changes.

CONFIDENTIALITY

Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to five years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information. In accordance with Title 13, no estimates are published that would disclose the operations of an individual firm.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

DISCLOSURE LIMITATION

A disclosure of data occurs when an individual can use published statistical information to identify either an individual or firm that has provided information under a pledge of confidentiality. Disclosure limitation is the process used to protect the confidentiality of the survey data provided by an individual or firm. Using disclosure limitation procedures, the Census Bureau modifies or removes the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual or business, the Census Bureau has taken steps to disguise or suppress the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of data.

UNPUBLISHED ESTIMATES

Additional statistics, such as estimates for some kinds of business not separately shown in this report, are produced as a byproduct of the regularly published statistics. These additional estimates have not been included in this publication because of high sampling variability, poor

response rates, or other factors that result in their failure to meet Census Bureau standards for publication. The Census Bureau, upon written request, will release such figures for individual use, though not for publication. It should be noted that some unpublished estimates can be derived directly from this report by subtracting published estimates from their respective totals. However, the figures obtained by such subtraction would be subject to the poor response rates or high sampling variability described previously for unpublished kinds of business.

Individuals who use Service Annual Survey estimates to create new estimates should cite the Census Bureau as the source of only the original estimates.

SAMPLE DESIGN AND ESTIMATION PROCEDURES

Introduction

A new sample was introduced with the 1999 Service Annual Survey (SAS). The new sample was designed to produce estimates based on the North American Industry Classification System (NAICS) and replaces the sample designed to produce estimates based on the Standard Industrial Classification (SIC) system. This section describes the design, selection, and estimation procedures for the new sample. For descriptions of prior samples, see the *Service Annual Survey* publications for years 1998 and earlier.

Sampling Frame

The sampling frame for the Service Annual Survey (SAS) has two types of sampling units represented -- Employer Identification Numbers (EINs) and large, multiple-establishment firms. Both sampling units represent clusters of one or more establishments owned or controlled by the same firm. The information used to create these sampling units was extracted from data collected as part of the 1997 Economic Census and from establishment records contained on the Census Bureau's Business Register as updated to June 1999. The next few paragraphs give details about the Business Register; the distinction between firms, EINs, and establishments; and the construction of the sampling units used in the sample selection. Though important, they are not essential to understanding the basic sample design and readers may continue to the **Stratification, Sampling Rates, and Allocation** section.

The Business Register is a multi-relational database that contains a record for each known establishment that is located in the United States or one of its territories and has employees. An *establishment* is a single physical location where business transactions take place and for which payroll and employment records are kept. Groups of one or more establishments under common ownership or control are *firms*. A *singleunit* firm owns or operates only one establishment. A *multiunit* firm owns or operates two or more establishments. The treatment of establishments on the Business Register differs according to whether the establishment is part of a singleunit or multiunit firm. In particular, the structure of an establishment’s primary identifier on the Business Register differs according to whether the establishment is part of a singleunit or multiunit firm.

A singleunit firm's primary identifier is its Employer Identification Number (EIN). The Internal Revenue Service (IRS) issues the EIN and the firm uses it as an identifier to report social security payments for its employees under the Federal Insurance Contributions Act (FICA). The same act requires all employer firms to use EINs. Each employer firm is associated with at least one EIN and only one firm can use a given EIN. Because a singleunit firm has only one establishment, there is a one-to-one relationship between the firm and the EIN. Thus the firm, the EIN, and the establishment all reference the same physical location and all three terms can be used interchangeably and unambiguously when referring to a single establishment firm.

For multiunit firms however, a different structure connects the firm with its establishments via the EIN. Essentially a multiunit firm is associated with a cluster of one or more EINs and EINs are associated with one or more establishments. A multiunit firm consists of at least two establishments. Each firm is associated with at least one EIN and only one firm can use a given EIN. However, one firm may have several EINs. Similarly, there is a one-to-many relationship between EINs and establishments. Each EIN can be associated with many establishments but each establishment is associated with only one EIN. Because of the possibility of one-to-many relationships, we must distinguish between the firm, its EINs, and its establishments. The multiunit firm that owns or controls a particular establishment is identified on the Business Register by way of the establishment's primary identifier.

The primary identifier of an establishment owned by a multiunit firm consists of a unique combination of an alpha number and a plant number. The alpha number identifies the multiunit firm; and the plant number identifies a particular establishment within that firm. All establishments owned or controlled by the same multiunit firm have the same alpha number. Different multiunit firms have different alpha numbers and different establishments within the same multiunit firm have different plant numbers. The Census Bureau assigns both the alpha number to the multiunit firm and plant numbers to the corresponding establishments based on the results of the quinquennial economic censuses and the annual Company Organization Survey.

To create the sampling frame, we extract the records for all establishments classified in selected service sectors as defined by the 1997 North American Industry Classification System. For these establishments, we extract revenue, payroll, employment, inventory, name and address information, as well as primary identifiers and, for establishments owned by multiunit firms, associated EINs. To create the sampling units for multiunit firms, we aggregate the economic data of the establishments owned by these firms to an EIN level by tabulating the establishment data for all service establishments associated with the same EIN. Similarly, we aggregate the data to a multiunit firm level by tabulating the establishment data for all service establishments associated with the same alpha number. No aggregation is necessary to put singleunit establishment information on an EIN basis or a firm basis. Thus, the sampling units created for singleunit firms simultaneously represent establishment, EIN, and firm information. In summary, the sampling frame is a complex amalgam of establishments, EINs, and firms.

Stratification, Sampling Rates, and Allocation

The primary stratification of the frame is by kind-of-business group based on the detail required

for SAS publications. We further stratify the sampling units within kind-of-business group (substratify) by a measure of size related to their annual revenue. Sampling units expected to have a large effect on the sampling error of the estimates are selected "with certainty." This means they are sure to be selected and will represent only themselves (i.e., have a selection probability of one and a sampling weight of one). Within each kind-of-business stratum we determined a substratum boundary (or cutoff) that divides the certainty units from the noncertainty units. We based these cutoffs on a statistical analysis of data from the 1997 Census of Service Industries. Accordingly, these values are on a 1997 revenue basis. We also used this analysis to determine the number of size substrata for each kind-of-business stratum and to set preliminary sampling rates needed to achieve specified sampling variability objectives on revenue estimates for different kind-of-business groups. The size substrata and sampling rates were later updated through an analysis of the sampling frame.

Sample Selection

The first step in the sample selection identified certainty firms. If a firm was selected with certainty and had more than one establishment at the time of sampling, any new establishments that the firm acquires, even if under new or different EINs, are included in the sample with certainty. However, if a singleunit firm was selected with certainty, only future establishments associated with that firm's originally-selected EIN are included in the sample with certainty; any new EINs that might later be associated with that firm are subjected to sampling through the quarterly birth-selection procedure (see **Sample Maintenance**).

All firms not selected with certainty were subjected to sampling on an EIN basis. If a firm had more than one EIN, each of its EINs was treated as a separate sampling unit. To be eligible for the initial sampling, an EIN used by a singleunit firm had to have nonzero payroll in 1998. EINs used by multiunit firms were required to have nonzero payroll in 1997. The EINs were stratified according to their major kind of business and their estimated revenue (on a 1997 basis). Within each noncertainty stratum, a simple random sample of EINs was selected. The maximum sampling weight for an EIN selected for SAS was 1,000.

Sample Maintenance

Periodically, we update the SAS sample to represent EINs issued since the initial sample selection. These new EINs, called births, are EINs recently assigned by the IRS, on the latest available IRS mailing list for FICA taxpayers, and assigned a kind-of-business classification (if possible) by the Social Security Administration (SSA).

EIN births are sampled on a quarterly basis (in November of the survey year and in February, May, and August of the year following the survey year) using a two-phase selection procedure. To be eligible for selection, a birth must either have no kind-of-business classification or be classified in a kind of business within the scope of SAS, the Annual Retail Trade Survey (ARTS), or the Annual Trade Survey (ATS), and it must meet certain criteria regarding its number of paid employees or quarterly payroll. In the first phase, births are stratified by kind of business and a measure of size based on expected employment or quarterly payroll. A relatively

large sample is drawn and canvassed to obtain a more reliable measure of size, consisting of revenues in two recent months, and a new or more detailed kind-of-business classification.

Using this more reliable information, the selected births from the first phase are subjected to probability proportional-to-size sampling with overall probabilities equivalent to those used in drawing the initial SAS, ARTS, and ATS samples from the June 1999 Business Register. Because of the time it takes for a new employer firm to acquire an EIN from the IRS, and because of the time needed to accomplish the two-phase birth-selection procedure, births are added to the samples approximately nine months after they begin operation.

For SAS, EIN births that are selected in the quarterly birth-selection procedure in November of the survey year are included in the initial mailing of the SAS questionnaires in January of the following year. To better represent all EIN births in the survey year, and specifically to account for the time it takes to identify and select new EINs, we add births to the sample that are selected in February, May, and August the year following the survey year. We mail survey forms to these births in June and August to supplement the initial survey mailing.

To be eligible for the sample canvass and tabulation in a given year, a noncertainty EIN must meet both of the following requirements:

- It must be on the latest available IRS mailing list for FICA taxpayers from the previous quarter.

- It must have been selected from the Business Register in either the initial sampling or during the quarterly birth-selection procedure.

EINs selected into the sample with certainty are not dropped from canvass and tabulation if they are no longer on the IRS mailing list. Rather, the firm that used the EIN is contacted, and if a successor EIN is found, it is added to the survey. For both inactive and reactivated EINs, data are tabulated for only the portion of the survey year that these EINs reported payroll to the IRS.

If a selected EIN ceases to be an employer, it becomes inactive. An inactive EIN is not mailed if it becomes inactive prior to the initial mailout of the survey year. An inactive EIN that resumes its employer status during the survey year is reactivated and mailed during the initial mailing (if active at the time) or as part of one of the two supplemental mailings.

Estimation Procedures for Annual Totals

Estimates in this report for Accommodation and Food Services industries are based on results from the Annual Retail Trade Survey (ARTS). All other service estimates in this report are based on results from the Service Annual Survey (SAS). All selected firms and EINs are asked to report data for the previous year. (Two years of data are requested in the year in which a new sample is introduced.) Estimates are computed as the sum of weighted data (reported and imputed) for all selected sampling units that meet the tabulation criteria given in the **Sample Maintenance** section. The weight for a given sampling unit is the reciprocal of its probability of

selection into the sample.

Estimates for some service industries presented in this report have been adjusted to results of the 1997 Economic Census. The NAICS industries for which we did not adjust estimates to results of the 1997 Economic Census are:

Newspaper, Periodical, Book, and Database Publishers (5111);
Sound Recording Industries (5122);
Cable Networks and Program Distribution (5132);
Paging (513321);
Cellular and Other Wireless Telecommunications (513322);
Other Information Services (514199);
Securities, Commodity Contracts, and Other Financial Investments and Related Activities (523);
Consumer Electronics and Appliances Rental (532210);
Title Abstract and Settlement Offices (541191);
Payroll Services (541214);
Computer Systems Design Services (541512);
Computer Facilities Management Services (541513);
Environmental Consulting Services (541620);
Translation and Interpretation Services (541930);
Private Mail Centers (561431);
Convention and Visitors Bureaus (561591);
Investigation Services (561611);
Security Guards and Patrol Services (561622);
Other Services to Buildings and Dwellings (561790);
Packaging and Labeling Services (561910);
Waste Treatment and Disposal (5622);
Remediation Services (562910);
Materials Recovery Facilities (562920);
All Other Miscellaneous Waste Management Services (562998);
Ambulance Services (621910);
Other Community Housing Services (624229);
Musical Groups and Artists (711130);
Other Performing Arts Companies (711190);
Marinas (713930);
Other Electronic and Precision Equipment Repair and Maintenance (811219);
Grantmaking Foundations (813211).

RELIABILITY OF THE ESTIMATES

The total error of an estimate based on a sample survey is the difference between the estimate and the true population value that it estimates. This error may be considered to be comprised of sampling error and nonsampling error. Sampling error is the difference between the estimate and the result that would be obtained from a complete census conducted under the same survey conditions. This error occurs because characteristics differ among sampling units and because

only a subset of the entire population is measured in a sample survey. Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate. The accuracy of a survey result may be affected by these two types of errors.

Sampling and nonsampling errors are often measured by the quantities, bias and variance. The *bias* of an estimator of an unknown population value is the difference, averaged over all possible samples of the same size and design, between the estimator and the unknown population value. Any systematic error, or inaccuracy that affects all samples of a specified design in a similar way, may bias the resulting estimates. The *variance* of an estimator is the squared difference, averaged over all possible samples of the same size and design, between the estimator and its average value.

Descriptions of sampling and nonsampling errors for the Service Annual Survey are provided in the following sections.

Sampling Error

Because the estimates are based on a sample, exact agreement with results that would be obtained from a complete enumeration of firms represented on the sampling frame using the same enumeration procedures is not expected. However, because each firm on the sampling frame has a known probability of being selected into the sample, it is possible to estimate the sampling variability of the survey estimates.

The particular sample used in this survey is one of a large number of samples of the same size that could have been selected using the same design. If all possible samples had been surveyed under the same conditions, an estimate of an unknown population value could have been obtained from each sample. These samples give rise to a distribution of estimates for the unknown population value. A statistical measure of the variability among these estimates is the standard error, which can be approximated from any one sample. The *standard error* is defined as the square root of the variance. The *coefficient of variation* (or relative standard error) of an estimator is the standard error of the estimator divided by the estimator. Note that measures of sampling variability, such as the standard error or coefficient of variation, are estimated from the sample and are also subject to sampling variability. (Technically, we should refer to the *estimated* standard error or the *estimated* coefficient of variation of an estimator. However, for the sake of brevity we have omitted this detail.) It is important to note that the standard error and coefficient of variation only measure sampling variability. They do not measure any systematic biases in the estimates. Coefficients of variation for 1999 and 2000 services estimates of total and e-commerce revenue are provided in Table 4A. These coefficients of variation are based on the 2000 SAS survey and, where appropriate, have been adjusted using results of the 1997 Economic Census. (All coefficients of variation are expressed as percents.) The Census Bureau recommends that individuals using Service Annual Survey estimates incorporate this information into their analyses, as sampling error could affect the conclusions drawn from the estimates.

The estimate from a particular sample and the standard error associated with the estimate can be used to construct a confidence interval. A *confidence interval* is a range about a given estimator

that has a specified probability of containing the result of a complete enumeration. Associated with each interval is a percentage of confidence, which is interpreted as follows. If, for each possible sample, an estimate of an unknown population value and its approximate standard error were obtained, then:

1. For approximately 90 percent of the possible samples, the interval from 1.65 standard errors below to 1.65 standard errors above the estimate would include the result of a complete enumeration.

2. For approximately 95 percent of the possible samples, the interval from two standard errors below to two standard errors above the estimate would include the result of a complete enumeration.

To illustrate the computation of a confidence interval for an estimate of total revenue, assume that an estimate of total revenue is $10,750 million and the coefficient of variation for this estimate is 1.8 percent, or 0.018. First obtain the standard error of the estimate by multiplying the total revenue estimate by its coefficient of variation. For this example, multiply $10,750 million by 0.018. This yields a standard error of $193.5 million. The upper and lower bounds of the 90-percent confidence interval are computed as $10,750 million plus or minus 1.65 times $193.5 million. Consequently, the 90-percent confidence interval is $10,431 million to $11,069 million. If corresponding confidence intervals were constructed for all possible samples of the same size and design, approximately 9 out of 10 (90 percent) of these intervals would contain the result obtained from a complete enumeration.

Nonsampling Error

Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate and may also occur in censuses. It is often helpful to think of nonsampling error as arising from deficiencies or mistakes at some point in the survey process. In the Service Annual Survey, nonsampling error can be attributed to many sources: inability to obtain information about all units in the sample; response errors; differences in the interpretation of the questions; mistakes in coding or keying the data obtained; and other errors of collection, response, coverage, and processing. Additional nonsampling error may have been introduced by the method used to adjust the survey estimates using results of the 1997 Economic Census. Although no direct measurement of the potential biases due to nonsampling error has been obtained, precautionary steps were taken in all phases of the collection, processing, and tabulation of the data in an effort to minimize their influence. The Census Bureau recommends that individuals using Service Annual Survey estimates incorporate this information into their analyses, as nonsampling error could affect the conclusions drawn from the estimates.

A potential source of bias in the estimates is nonresponse. Nonresponse is defined as the inability to obtain all the intended measurements or responses about all selected units. Two types of nonresponse are often distinguished. *Unit nonresponse* is used to describe the inability to obtain any of the substantive measurements about a sampled unit. In most cases of unit nonresponse, the questionnaire was never returned to the Census Bureau, after several attempts to

elicit a response. *Item nonresponse* occurs either when a question is unanswered or the response to the question fails computer or analyst edits.

For both unit and item nonresponse, a missing value is replaced by a predicted value obtained from an appropriate model for nonresponse. This procedure is called *imputation* and uses survey data and administrative records as input. For SAS, imputed revenue amounts to about 12.0 percent of the total revenue estimate and about 29.0 percent of the total e-commerce revenue estimate.